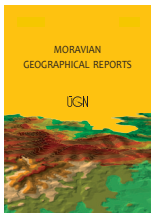

MORAVIAN GEOGRAPHICAL REPORTS

Institute of Geonics, The Czech Academy of Sciences

journal homepage: http://www.geonika.cz/mgr.html

doi: 10.1515/mgr-2016-0008

DE GRUYTER OPEN

Delineating zones to increase geographical detail in individual response data files: An application to the Spanish 2011 Census of population

Lucas Martínez-Bernabeu [a *], José Manuel Casado-Díaz [b]

Abstract

Due to confidentiality considerations, the microdata available from the 2011 Spanish Census have been codified at a provincial (NUTS 3) level except when the municipal (LAU 2) population exceeds 20,000 inhabitants (a requirement that is met by less than 5% of all municipalities). For the remainder of the municipalities within a given province, information is only provided for their classification in wide population intervals. These limitations, hampering territorially-focused socio-economic analyses, and more specifically, those related to the labour market, are observed in many other countries. This article proposes and demonstrates an automatic procedure aimed at delineating a set of areas that meet such population requirements and that may be used to re-codify the geographic reference in these cases, thereby increasing the territorial detail at which individual information is available. The method aggregates municipalities into clusters based on the optimisation of a relevant objective function subject to a number of statistical constraints, and is implemented using evolutionary computation techniques. Clusters are defined to fit outer boundaries at the level of labour market areas.

Key words: *labour market areas, census, microdata, regionalisation, clustering, evolutionary computation, Spain*

Article history: *Received 30 September 2015; Accepted 10 December 2015; Published 30 June 2016*

1. Introduction

As in many other countries, microdata derived from a Census of Population are a very rich source of information for the analysis of socio-economic phenomena in Spain. One of the potential applications of this dataset is conducting labour market analyses at very detailed territorial levels, something that is not feasible when using other, more frequently updated, sources of information such as a Labour Force Survey, due to their sampling limitations. When attempting to conduct analyses based on Census microdata, however, researchers and other potential users are faced with the fact that (e.g., in the Spanish case) the geographic reference provided for the majority of indicators appears at a provincial level (NUTS 4), with information on the reference municipality (LAU 2) available only when the population of said municipality exceeds some threshold (e.g., 20,000 inhabitants in Spain). In the remainder of the cases, the information is aggregated into four groups of municipalities[1] for each province, mainly due to confidentiality and sampling constraints. In these cases, apart from the province code, there is only information provided on the population category to which the municipality of residence belongs (e.g. the municipality of residence belongs to province x and is in the range 2,001– 5,000 inhabitants). Since about 95% of the 8,116 Spanish municipalities have less than 20,000 inhabitants, this characteristic of the microdata set results in the loss of a large amount of potentially useful information[2]. These restrictions apply to seven territorial variables: place of residence, place of birth, previous place of residence, place of residence one year ago, place of residence ten years ago, place of second residence and place of work.

The motivation for this article is therefore practical. It seeks to produce a geography that allows the re-codification

[a] International Economics Institute, University of Alicante, Spain (*corresponding author: L. Martínez-Bernabeu, e-mail: *lucas.martinez@ua.es*)

[b] International Economics Institute and Department of Applied Economic Analysis, University of Alicante, Spain

[1] Group 1: municipalities with less than 2,001 inhabitants; group 2: between 2,001 and 5,000; group 3: between 5,001 and 10,000; and group 4: between 10,001 and 20,000.

[2] This fact is discussed in detail in section 5.

of the territorial variables in the Census microdata file in order to regain as much spatial information as possible, so that the currently vague reference to population intervals for less-populated municipalities may be substituted by a reference to specific clusters of municipalities. Such clusters are designed to meet the statistical constraints imposed by the National Institute of Statistics, which in this case refer only to a minimum population threshold, and to nest in the upper-level geography of labour market areas (LMAs, a set of functional areas which was defined in other research: Martínez-Bernabeu et al., 2016)[3]. Some of the LMAs have a reasonable level of spatial resolution to be used as the geographic reference in order to re-codify the microdata, but many of them could be further sub-divided if the only requirement to be fulfilled is that of having a minimum population size of 20,000. Therefore, to further increase territorial detail, this article aims at sub-dividing these LMAs into so-called "municipality clusters", with populations over 20,000 inhabitants for which a minimum self-containment level is not required. As in the case of LMAs, this regionalisation is characterised by an exhaustive coverage of the entire territory under consideration, not allowing overlapping between the resulting areas and enforcing contiguity between the municipalities making up each cluster.

Fulfilling the objective of this article, the subdivision of the Spanish LMAs into their constituting clusters of municipalities to increase territorial detail in the Census 2011 individual data sets, involved the definition of a new procedure based on evolutionary computation. Such a procedure has been tailored to fit the specific characteristics of the problem, since despite being guided by the commuting links between the municipalities and the interaction between clusters in these same terms, the process of delineation of clusters is quite different from the identification of LMAs. Thus, while the aim of the definition of the Spanish LMAs was to maximise the internal interaction between the constituting municipalities within each LMA subject to the fulfilment of both a minimum population condition and a trade-off rule between self-containment and area (Martínez-Bernabeu et al., 2016), the aim of this work is to maximize the number of clusters identified so that each LMA is sub-divided into as many clusters as possible, each of which must exceed the minimum population threshold.

The remainder of this article is organized as follows: Section 2 provides the background for the analysis and in Section 3 the different elements making up the problem are described in detail. The latter include the problem formulation as an optimization procedure subject to certain constraints, guided by a fitness function based on an interaction (in terms of travel-to-work) index. In Section 4, the evolutionary algorithm (structure, chromosome representation, operators and configuration/parameters) used is described in detail. The resulting set of 931 municipality clusters is presented and discussed in Section 5. Finally, Section 6 offers some conclusions.

2. Background

The problem addressed in this article -- grouping a set of elements with an associated size (or cost) into as many disjoint groups subject to reach a minimum size as possible – is a specific case of the more general Set Partitioning (SP) problem (Balas and Padberg, 1976). In the SP problem, the input is a finite set of elements, U, called a universe, and a set of possible subsets of the universe, S, each with an associated cost. The task is to find the partition P (i.e. a subset of S so that all sets in P are pairwise disjoint and the union of P is equal to the universe) with minimum total cost, calculated as the sum of the costs of each subset in the partition. This is a complex problem (NP-complete) having numerous real-life applications, e.g. airline crew scheduling (Barnhart et al., 2003) and vehicle routing (Toth and Vigo, 2001). Most of the applications of the SP problem solve it through integer programming for small instances and approximation algorithms for instances that become computationally intractable through exact methods (Laporte, 1992). Other forms of optimisation methods, particularly genetic algorithms, have also been successfully used (e.g. Levine, 1996), and are particularly useful when facing large instances in which linear relaxations and approximations for the integer programming approach do not suffice to make them computationally tractable.

This article focuses on a specific instance of this problem. Such an instance has some peculiarities compared with the general SP problem: its objective is to maximise the value of the partition instead of minimising its cost; and, more importantly, the number of possible subsets of the universe, S, is not an input to the problem (i.e. it is unknown a priori). Instead of generating a huge set of possible subsets in a first step and then solving the associated SP problem, these approaches solve both problems simultaneously by applying a stochastic optimisation method that performs a randomised search over possible partitions.

One example of such a family of applications is the delineation of Census "output areas" (OAs). This consists of the grouping of a given set of spatial building blocks into subsets which are argued to be appropriate for the publication and the integration of different datasets derived from a Census of Population. In the case of the UK, OAs of the 2011 Census were defined[4] for England and Wales using the "automated zoning procedure" (AZP) originally designed by Openshaw (1977a and b) and further refined by Openshaw and Rao (1995). This procedure departs from a possible regionalisation of OAs (the definition of such areas specifically produced for the previous Census), and iteratively re-allocates building blocks, chosen at random, between OAs, accepting one specific re-allocation if it improves the design criteria and otherwise rejecting it, until no more positive re-allocations are found after a certain number of iterations. In the case of the OAs (Martin et al., 2001), such criteria included a constraint in terms of minimum population and three objectives to be optimised (with each given the same weight): a target population criterion (minimising the sum

[3] LMAs were defined using a variation of the so-called GEA method (Martínez-Bernabeu, et al., 2012). The output of such a process was the partition of the 8,116 Spanish municipalities into a total of 260 non-overlapping LMAs made up of one or more contiguous municipalities, with each LMA having a population exceeding 20,000 inhabitants and a self-containment of over 70% (i.e., at least 70% of local jobs are taken by residents of the area, and at least 70% of the residents work locally).

[4] On the occasion of the new 2011 Census of Population, maintenance of the existing set of OAs was preferred to the complete re-design of this set of zones. This involved splitting, merging or re-designing a small sub-set of existing OAs, a process that was based on AZP but that required more manual intervention as compared to the original delineation process conducted in 2001 (Cockings et al., 2011).

of the squared differences between OA populations and the specified target population within each administrative area), the (within zone) social homogeneity (measured as the intra-area correlation in terms of dwelling type and tenure); and morphological compactness (which implied minimising the squared perimeter divided by area). The stochastic nature of this procedure allows for an automatic search over the possible regionalisations without the need to implement complex heuristics, but it does have some handicaps. First, it only considers single building blocks re-allocations and only accepts them if they improve the design criteria. Therefore, it does not allow for an exhaustive search of the solutions' space and will get trapped in local maxima if the problem is not trivial. Second, it does not allow changes to be made to the initial number of OAs, which remains fixed as the number of regions of the initial solution or by user input if no initial regionalisation was provided, and that is a problem when there is no a priori knowledge regarding the appropriate number of regions.

A different group of SP problems that has connections with the one on which this article focuses, is that of tackling with the definition of LMAs: areas aimed at capturing the local dimension of labour markets understood as the spaces where local supply and demand for labour meet. Ideally, each LMA should be characterised by being externally self-contained in terms of commuting to work (i.e., there are few commuters travelling between different LMAs), and by being internally integrated in those same terms (i.e., the ideal LMA should consist of basic building blocks among which daily commuting flows are abundant). Although international experience is quite extensive (see, for example, Casado-Díaz and Coombes, 2011), only a limited number of authors have dealt with the problem of delineation of LMAs as a SP problem[5]. Authors who have addressed this issue include Flórez-Revuelta et al. (2008), Farmer and Fotheringham (2011), Fusco and Caglioni (2011), Martínez-Bernabeu et al. (2012), Chakraborty et al. (2013) and Alonso et al. (2015).

Flórez-Revuelta et al. (2008) proposed a grouping evolutionary algorithm (a general-purpose optimisation technique used in Artificial Intelligence, with genetic operators specifically designed to fit grouping problems) in order to optimise a fitness function that measures the interaction within LMAs, subject to reach certain minimum self-containment and population thresholds. Their fitness function is based on the interaction index (originally proposed by Smart, 1974) that is used to define the official Travel-to-Work Areas (TTWAs) in the UK (Coombes et al., 1986; Coombes and Bond, 2008; ONS, 2015) and the Sistemi Locali del Lavoro in Italy (ISTAT, 1997; 2005; 2014), their local version of LMAs. Martínez-Bernabeu et al. (2012) further improved upon the work by Flórez-Revuelta et al. (2008) by designing renovated search operators that allow for higher quality results and a reduction in computational costs. Alonso et al. (2015) propose and exemplify a delineation scheme based on these grouping evolutionary algorithms.

The work by Farmer and Fotheringham (2011) and Fusco and Caglioni (2011) use a different objective function, the modularity quality index. This function, borrowed from Newman and Girvan (2004), was originally developed for the detection of (social) communities in networks. It accumulates the difference between the interaction links within each community and their expected value in a network having the same nodes but with uniformly distributed flows (the null model). The use of the modularity function has been criticised in the context of community detection (Fortunato and Barthelemy, 2007; Lancichinetti and Fortunato, 2011), since it is unable to identify communities (that are obvious to the human eye) when the number of nodes vary sufficiently between different communities (or, in the LMA context, when large variations between the actual LMAs are observed in population terms). Moreover, the expected interaction value in the null model increases with the size of the territory under analysis, while the actual LMAs for a given region should not depend on whether or not some other unrelated regions are included in the analysis. These drawbacks of the modularity function lead to our preference for the interaction function of Flórez-Revuelta et al. (2008), as well as their general methodology, which has been found to produce better results than the widely- applied TTWAs method, in terms of the number of identified LMAs and cohesion values for the same levels of minimum self-containment, while the works based on modularity have not been compared with alternative approaches.

Since this article focuses on the problem of identifying subsets of municipalities within each LMA, it was considered important to retain the assessment of the commuting links at a cluster level as part of the delineation process (and this is a type of variable that is not considered in the OAs definition process, which is based on the attributes of the building blocks and not on the functional relationships observable between them). Moreover, AZP suffer some technical inconveniences that have been outlined above. This led us to favour the adaptation of a different grouping algorithm (GEA: see below) in order to tackle this specific problem instead of adopting any of the other obvious alternatives. This process has involved defining a fitness function, constraints and a set of operators adapted to this specific grouping problem.

3. Problem statement

As stated in the previous sections, the problem consists of the within-LMA grouping of basic spatial units (BSU), in this case municipalities, into as many geographically continuous clusters of municipalities with a minimum size of 20,000 inhabitants as possible. Thus, the number of identified clusters is the main objective to be maximised.

We also introduce the maximisation of the interaction between municipalities within each area as a secondary objective. That is, we shall always prefer producing (continuous) groupings consisting of $(n + 1)$ clusters over groupings of n clusters, but when facing two alternative groupings with the same number of clusters, we shall prefer the one with the higher inner interaction. Thus, the defined clusters shall be as connected as possible, avoiding the identification of clusters composed of BSUs that are not linked by commuting flows whenever possible.

3.1 Problem formulation

Let $U = \{1, 2, \ldots N\}$ be a set of $N = |U|$ BSUs (the LMA to be divided into clusters of municipalities); T, the matrix of commuting flows, so that T_{ij} is the number of commuters

[5] The methods applied can be more often characterised as greedy: they use one or more heuristics that quickly produce a reasonable but sub-optimal regionalisation, through methods that are not based on a fitness function to be maximised and therefore cannot be characterised as optimisation procedures.

from BSU i to BSU j; and P, the vector of populations, so that P_i is the population of BSU i. The objective is to obtain the set of clusters C = {C1, C2, ... CK} that maximises the fitness function f(T,P,C), described in section 3.2, subject to

- C being a partition of U (i.e., $C_i \neq \varnothing \ \forall \ C_i \in$ C; $\cup_{i = 1...K} C_i = U$, and $C_i \cap$ Cj $= \varnothing \ \forall \ C_i, C_j \in$ C, $i \neq j$),
- $\sum_{x \in C_i} P_x \geq 20{,}000 \ \forall$ Ci $\in$ C, and
- each cluster C_i being geographically continuous.

3.2 Commuting interaction index

To assess the degree of commuting interaction between a pair of clusters or BSUs, we use the interaction index proposed by Flórez-Revuelta et al. (2008), a generalisation of the index used in the TTWA method (Coombes et al., 1986). This indicator takes into account the commuting flows in both directions as well as the relative size of both areas to weight the flows between them. Thus, the flows between small interdependent areas are not eclipsed by the flows between larger areas. Let the interaction index between two clusters II(Ci,Cj) be defined as:

$$\mathrm{II}(C_i, C_j) = \mathrm{II}(C_j, C_i) = \frac{T(C_i, C_j)^2}{R_i \cdot J_j} + \frac{T(C_j, C_i)^2}{R_j \cdot J_i} \qquad (1)$$

in which $T(C_i,C_j)$ is the number of commuters from any of the BSUs in C_i to any of the BSUs in C_j; $R_k = T(C_k, U)$ is the total number of workers residing in C_k; and $J_k = T(U,C_k)$ is the total number of jobs in C_k.

3.3 Fitness function

The main objective of maximising the number of identified clusters may be directly represented by the number of identified clusters. The interaction within clusters may be measured with the same fitness function used in Flórez-Revuelta et al. (2008):

$$g(C) = \sum_{i \in U} \mathrm{II}(\{i\}, C^{(i)}) \qquad (2)$$

in which $C^{(i)}$ represents the cluster to which BSU i belongs minus the own BSU i, and $\{i\}$ represents the cluster formed by i alone. This function accumulates the interaction value between (a) each BSU i and (b) the aggregation of the rest of BSUs in the cluster which that specific BSU i is a part of (excluding T_{ii}).

In order to include the interaction value in the fitness function as a secondary objective, to the number of identified areas we add the average global interaction per BSU, with values in the range [0, 1] (that in practice are always close to 0). Thus, the secondary objective can never force the choice of a grouping of n areas over one of $(n + 1)$ areas, but different groupings of n areas will have different evaluations, depending on the associated interaction levels, and it will allow us to choose the one having more within-clusters interaction:

$$f(C) = \mathrm{card}(C) + \frac{g(C)}{N} \qquad (3)$$

4. Optimisation algorithm

We base our proposal on the grouping evolutionary algorithm (GEA) by Martínez-Bernabeu et al. (2012). This type of algorithm, within the family of genetic algorithms (Goldberg, 1989), is based on the principles of natural evolution and the selection of the fittest. Generally speaking, genetic algorithms are stochastic optimisation techniques, and the specific class of grouping genetic algorithms (Falkenauer, 1998) use tailored genetic operators working over an encoding that can represent groupings of elements, in this case clusters of municipalities within LMAs.

Departing from an initial population of solutions (called *individuals*), which are codified as numeric *chromosomes*, new solutions are created by combining the current individuals (as in sexual reproduction) and applying random changes (as in genetic mutations) to the chromosomes. Then, the new individuals are evaluated using a *fitness function* and some of them are chosen (using a *selection scheme* that favours solutions with better evaluations) to remain in the population for the next iteration (called *generation*) of the algorithm, until a certain *stop condition* is met. This is described in detail in the following subsections, where three forms of stochastic selection are used: random (i.e. uniform probability), probability proportional to the attraction (self-explanatory), and 3-way tournament[6] over a certain characteristic (attraction, size, etc.).

4.1 Structure of the optimisation algorithm

The structure of the GEA algorithm follows these steps:

1. Initialise population: Generate N_p valid solutions by taking the whole set of BSUs in U as mono-BSU clusters and apply the greedy heuristic SHA (described in section 4.3) over them;
2. Evaluate fitness and rank population;
3. Repeat until no improvement of the best solution is found for N_g generations:

3.1 Apply genetic operators until N_o new valid individuals are produced, as follows:

3.1.1 Select a parent from the current population with a probability proportional to the fitness ranking;

3.1.2 Randomly select an operator with uniform probability;

3.1.3 If the operator is the crossover, select a second, different parent with a probability proportional to the fitness ranking;

3.1.4 Create a new individual as a copy of the (first) parent;

3.1.5 Apply the selected operator to the new individual;

3.1.6 If the operator terminates successfully, the resulting individual is evaluated; otherwise its fitness will be set to 0 (invalid);

3.2 Rank individuals in the population by their fitness; and

3.3 From the current pool of previous and new individuals, select the N_p individuals that will stay in the population for the next generation, using selection by ranking with elitism for the best.

The N_p parameter defines the population size, the N_o parameter controls how many new individuals are generated in each generation, and the N_g parameter controls how many generations without further improvement will be performed before stopping the search. In our application we set $N_p = 25$ and $N_o = 10$ and $N_g = 5{,}000$.

[6] This is performed by randomly selecting three elements and then selecting the one with highest (or lowest) score in the given characteristic.

In contrast with regular genetic algorithms, the crossover and mutation operators are treated equally in a single stage, so that No mutations and no crossovers (or vice versa) could be applied in a given generation.

4.2 Chromosome representation

We use exactly the same representation as in Martínez-Bernabeu et al. (2012), referred to as group-number encoding: the chromosome of an individual is a vector of N integers (one for each BSU in U), so that the BSUs with the same integer value are allocated to the same cluster. This representation ensures that the solution is a partition and that the corresponding constraints are automatically met (that is, each BSU is assigned to one and only one cluster). The integer values on each chromosome are forced to follow an ascending order to avoid the possibility of having different representations for the same partition, so that for a partition of x clusters, the first BSU is always assigned to group 0, the following BSU allocated to a different group will be assigned group 1 (and so on), and the maximum group number will be $x - 1$.

4.3 Stochastic Hierarchical Agglomeration

For the creation of the initial population (step 1 in the GEA algorithm, see section 4.1), as well as for the reparation of the invalid clusters that may result from the crossover operator (described in section 4.4), we adapt the greedy heuristic presented in Martínez-Bernabeu et al. (2012), the Stochastic Hierarchical Agglomeration (SHA). This algorithm starts from a given partition of a set: one cluster per BSU in the case of step 1 of GEA, or the partition resulting from the crossover operator (that will normally include clusters with several BSUs). Then, it iteratively chooses a cluster with population lower than 20,000 and another adjacent area with low population, and merges them, repeating these steps until all of the clusters have at least 20,000 inhabitants. The exact procedure followed in this work is as follows:

1. Terminate successfully if all the clusters have at least 20,000 inhabitants;
2. Select a cluster G by 3-way tournament over the inverse of population;
3. Select a cluster H adjacent to G, with a probability proportional to the inverse of its population; and
4. Merge clusters G and H and go to step 2.

4.4 Grouping genetic operators

Martínez-Bernabeu et al. (2012) describe ten group-based genetic operators: one crossover and nine mutations designed to cover all general operations over what in mathematical terms are known as disjoint sets. In this study, we have used the crossover and only five of those mutation operators (M, I, E, D and N), adapted to the particular objectives of our specific grouping problem.

This has affected the attraction criteria between pairs of clusters: while the original operators use the commuting interaction index (eq. 1), to help maximise the main objective of LMA definition (interaction within LMAs), our variants use the inverse of the summation of both cluster's population, to contribute to the maximisation of the number of clusters identified, the main objective of the process:

$$a(C_x, C_y) = \frac{1}{\sum_{i \in C_x} P_i + \sum_{i \in C_y} P_i} \quad (4)$$

The following subsections describe the precise algorithms of each operator used in this work.

4.4.1 Crossover

This operator is based on the standard grouping crossover as described by Falkenauer (1998). A random selection of clusters from one parent is copied over the other, changing the codification so that none of the copied clusters share their code number with any of those already present in the recipient parent. The integrity of the copied clusters is maintained while the clusters of the other parent sharing BSUs with them can became invalid in terms of size or contiguity. Any non-continuous cluster is fragmented into (smaller) contiguous clusters. Then, the SHA procedure (see section 4.3) is applied to all individuals, so that invalid fragments of clusters are merged with adjacent clusters until they are all valid. This procedure can also modify the initially preserved clusters (those that absorb other invalid clusters):

1. Copy all of the information from the first parent into the child;
2. Randomly select a number r with uniform distribution between 1 and 66% of the amount of clusters in the second parent;
3. Randomly select r distinct clusters from the second parent and copy them into the child, changing the codification so that none of that clusters share their code with any cluster in the offspring;
4. Check each cluster and divide those clusters that are not continuous into their continuous parts;
5. Apply SHA over all of the clusters of the child (reparation of broken clusters from the first parent); and
6. Terminate successfully.

4.4.2 Mutation M: random re-allocations

This operator randomly selects *border*[7] BSUs with low interaction in their clusters, and attempts to re-allocate them to other adjacent clusters. This is the operator closer to the concept of standard mutation in general genetic algorithms:

1. Randomly select a number r between 1 and 2% of the total number of BSUs;
2. Repeat r times:

2.1 Choose a border BSU i with low attraction with its micro-area C_i by 3-way tournament;

2.2 Select a cluster C_j adjacent to i, with probability proportional to the attraction to I;

2.3 Re-allocation of i from C_i to C_j if both clusters continue being valid; and

3. If at least one effective re-allocation occurred, terminate successfully; otherwise terminate unsuccessfully.

4.4.3 Mutation I: inclusion into a cluster of adjacent BSUs

This operator attempts to increase the size of a cluster with a low population by absorbing some of the adjacent,

[7] A border BSU is one that is adjacent to at least one cluster other than the one that it is currently part of.

unnecessary[8] BSUs into the surrounding clusters with which it shares higher interaction (the opposite of the mutation *E*):

1. Select a cluster C_i with low population by 3-way tournament;
2. Randomly select a number r between 1 and 10% of the average number of BSUs per cluster;
3. Repeat r times:

   3.1 Select a BSU i adjacent to C_i and belonging to a cluster $C_j \neq C_i$ with a probability proportional to the attraction to C_i;

   3.2 Re-allocate i from C_j to C_i if both clusters continue being valid; and
4. If at least one effective re-allocation occurred, terminate successfully; otherwise terminate unsuccessfully.

4.4.4 Mutation E: exclusion of border BSUs with high external attraction from a cluster

This operator attempts to reduce the size of a large cluster by choosing some border BSUs with lower interaction with the rest of the cluster to which it is currently assigned and reassigning them to other related, adjacent clusters. Its process is inverse to that of mutation *I*:

1. Select a cluster C_i of high population by 3-way tournament;
2. Randomly select a number r between 1 and 20% of the amount of BSUs in C_i;
3. Repeat r times:

   3.1 Select a border BSU i from C_i having a low population by 3-way tournament;

   3.2 Select a cluster C_j adjacent to i with a probability proportional to the attraction to I;

   3.3 Re-allocate i from C_j to C_i if both clusters continue being valid; and
4. If at least one effective re-allocation occurred, terminate successfully; otherwise terminate unsuccessfully.

4.4.5 Mutation D: dismembering of a cluster and assignation of its constituent BSUs to the adjacent clusters

This operator uses the same mechanism as in mutation E, but finishes only when the cluster disappears. This operator will always reduce the number of clusters by one, worsening the fitness value, but this may allow that subsequent operations find a better solution and help the search process to escape from a local maximum:

1. Select a cluster C_i with low population by3-way tournament;
2. Repeat until there are no remaining BSUs in C_i:

   2.1 Select a border BSU i from C_i with low population by 3-way tournament;

   2.2 Select a cluster C_j adjacent to i with a probability proportional to the attraction to I;

   2.3 Re-allocate i from C_i to C_j if C_j continues being valid after the re-allocation, otherwise terminate unsuccessfully; and
3. Terminate successfully.

4.4.6 Mutation N: creation of a new cluster using a border BSU as seed

This operator chooses an unnecessary, border BSU in a cluster of low population, creates a new cluster from that BSU, and then tries to absorb other unnecessary, adjacent BSUs from surrounding clusters, until the new cluster reaches the minimum population or there are no more available BSUs to absorb:

1. Select a cluster C_i with a high population by 3-way tournament;
2. Select an unnecessary, border BSU i from C_i with a low population by 3-way tournament. If it cannot be found, terminate unsuccessfully;
3. Create a new cluster C_j conformed by I;
4. Repeat while population of C_j is smaller than 20,000:

   4.1 Select BSU k from the BSUs adjacent to C_j, with a probability proportional to the attraction to C_j;

   4.2 Re-allocate k from its cluster C_k to C_j if C_k continues being valid after the re-allocation, otherwise terminate unsuccessfully; and
5. Terminate successfully.

5. Results

Of the 260 LMAs which, according to the objective of this article, should be divided into clusters, 86 already had a population of less than 40,000 inhabitants, and therefore a subdivision was not possible. Thus, the grouping technique described in this paper was applied to the remaining 174 LMAs whose populations exceeded 40,000 inhabitants. Of these, 21 LMAs could not be divided because one of the BSUs concentrated most of the population and any grouping of the remaining BSUs could not reach the minimum of 20,000 inhabitants (10 cases), or because they were formed by only one BSU (2 cases: the cities of Ceuta and Melilla, in the north of Africa), or because the contiguity restriction did not allow for a proper division (9 cases). The remaining 153 LMAs were divided into 824 clusters (totalling 931 clusters with the undivided LMAs[9]). As expected, the LMAs that were sub-divided into a larger number of clusters are those centred in the largest metropolitan areas: Madrid (with 60 clusters), Barcelona (34), Valencia (32), Terrassa (32), Sevilla (22) and Bilbao (22).

To assess the extent to which the results increase and improve the territorial detail of the original reference geography of municipalities grouped in ranges of population, we have compared both regionalisations. Figure 1 depicts the geography that currently serves as a territorial reference in the conventionally distributed Census microdata, as described in Section 1. Such geography consists of the 402 municipalities whose population exceeds 20,000 inhabitants (coloured in dark blue), plus the within-province aggregation of the remaining municipalities into groups according to their population range (these groups have been coloured in blue shades according to the specific population group to which their municipalities belong[10]). The combination of both territorial references (large municipalities plus the within-

[8] An unnecessary BSU is one that can be re-allocated to another cluster without breaking the constraints of minimum population and contiguity.

[9] These clusters include one for which the population minimum is not reached: El Hierro (in the Canary Islands). This is a very specific case whose separate consideration is justified since it is the only populated island not reaching the minimum population threshold, despite being one of the territories with higher self-containment levels.

[10] It is noticeable that 28 of such population-range sub-provincial clusters have in fact less than 20,000 inhabitants.

province population groups) results in the 587 clusters that are presented in Figure 1. It is noticeable that in many provinces, one specific cluster (that consisting of the municipalities under 2001 inhabitants) covers most of the area, and that, in general, clusters based on population ranges are formed by fragmented parts among which distances may be very large.

On the other hand, Figure 2 shows the 931 clusters of municipalities obtained with our methodology, which as previously noted, has been applied within each of the LMAs defined in a previous article (the colour scale reflects the clusters' population levels). In this regionalisation, only 171 clusters are formed by a single municipality. Figure 3 focuses on a specific example: the province of

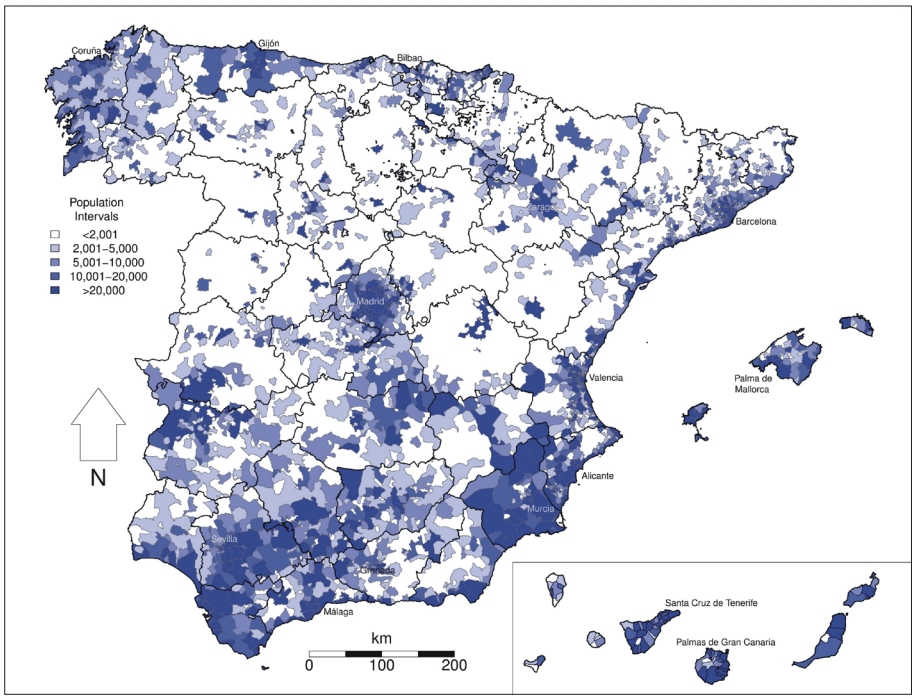

Fig. 1: Geographic reference currently in use in the Spanish Census microdata file 2011. Source: Authors' results based on data from the Spanish Census of Population 2011 (Instituto Nacional de Estadística, INE). Notes: Black lines mark provincial boundaries. Microdata are currently referenced to 587 regions (402 individual municipalities whose population exceeds 20,000 inhabitants – marked in the Figure with the darkest shade – plus within-province groupings of the remaining municipalities according to the population ranges depicted in the Figure's legend – within each province municipalities marked with the same colour belong to one cluster)

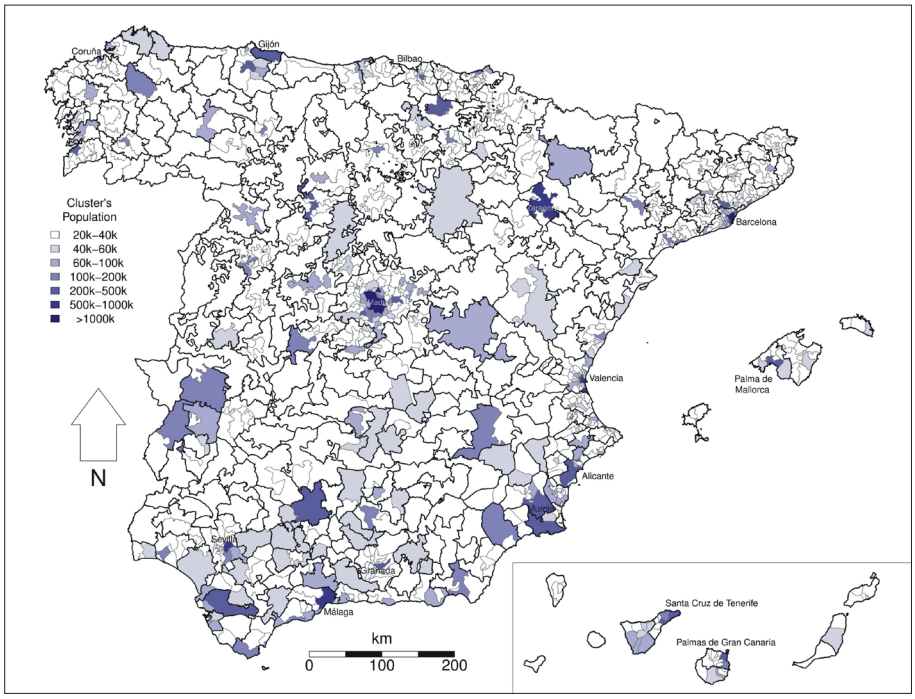

Fig. 2: The proposed geography. Source: Authors' results based on data from the Spanish Census of Population 2011 (Instituto Nacional de Estadística, INE). Notes: Black lines mark LMAs' boundaries. Grey lines mark clusters' boundaries (all clusters are formed by continuous municipalities). The colour scale characterises each cluster according to its population

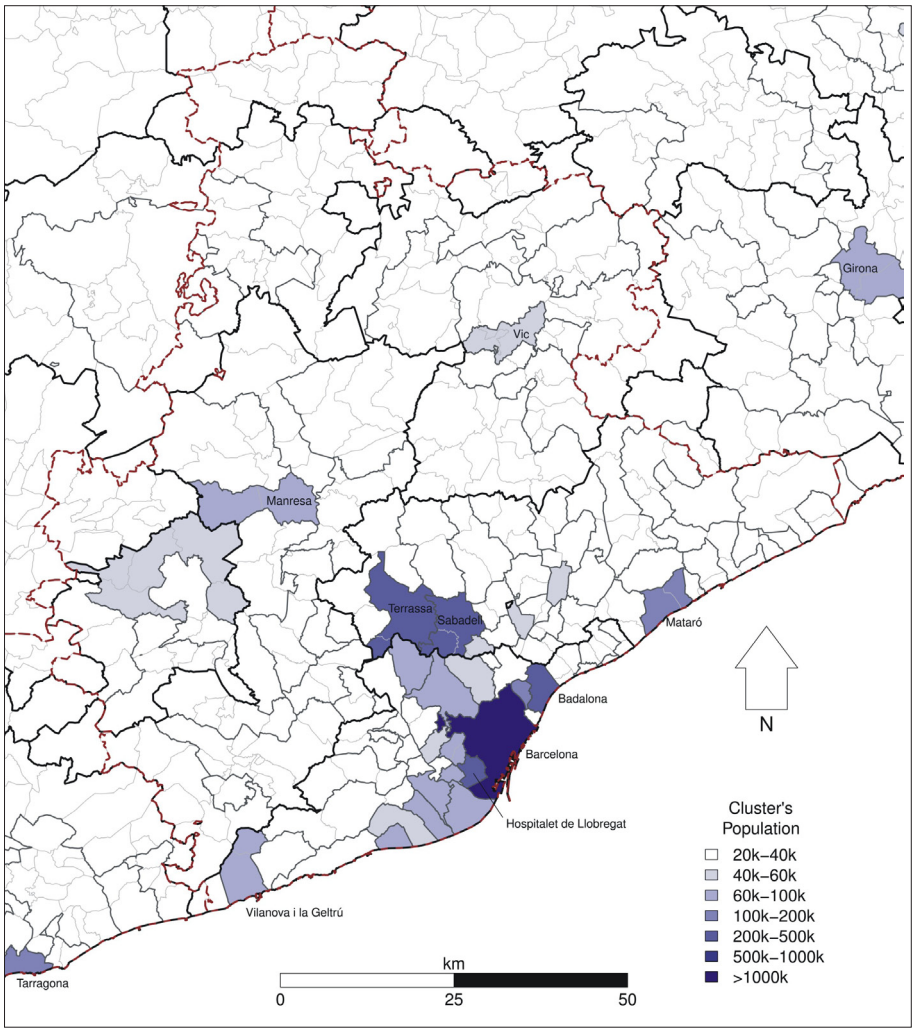

Fig. 3: The proposed geography. Detail of the province of Barcelona. Source: Authors' results based on data from the Spanish Census of Population 2011 (Instituto Nacional de Estadística, INE). Notes: Red line marks provincial boundaries. Black lines mark LMAs' boundaries. Darker grey lines mark clusters' boundaries (all clusters are formed by continuous municipalities). Municipality boundaries in light grey lines. The colour scale characterises each cluster according to its population

Barcelona. This figure illustrates how clusters are structured at a municipality level. Thus that figure shows the actual groupings of municipalities within each cluster of the approximately six LMAs that cover the province.

Morphologically, we see four relevant differences between the original regionalisation (Figure 1) and our results (Figures 2 and 3): (a) our proposal involves a great increase in the microdata territorial detail since it consists of almost 60% more regions, that are more comparable in terms of area; (b) specifically, our proposal divides the large clusters of municipalities that almost completely cover some of the Spanish inner provinces in the current geography into several clusters; (c) many of the municipalities that form a singleton cluster in the currently-used geography are grouped with other smaller municipalities in our proposal, although none of them becomes considerably large; and (d) all of the clusters in our proposal are contiguous. And, obviously given the design of the methodology applied, the clusters of our proposal honours the boundaries of the LMAs of this study.

To complete the description of the proposed geography and its comparison with that which is currently in use, Tables 1 and 2, respectively, depict the number of clusters by area and population intervals. Each table includes, for each regionalisation (the one currently used and the geography proposed here), one column with information for all clusters and a second column in which only the clusters that group at least two (2) municipalities are considered (this column has been labelled "> 1 municipalities").

As shown in both tables, the regionalisation resulting from the method applied in this article offers a much higher level of territorial detail. Thus, in Table 1, it is noticeable that while in the currently-used geography, 19 clusters have an area over 6,000 km^2, this threshold is not exceeded in any of the clusters included in our proposal, and only 6 clusters are over 4,000 km^2, so that microdata records can be referenced to smaller, more specific geographical places. In terms of population (Table 2), the proposed geography of clusters also involves a great increase in the level of detail. In this case, most of the gain (compared with the current geography) occurs in the range between 60,000 and 300,000 inhabitants.

Thus the current territorial division has 335 clusters in the range of 20,000 to 60,000 inhabitants (i.e. 57% of clusters), whereas in our proposal 793 clusters (85.2%) fall within that interval. On the other hand, the geography currently in use includes 209 clusters in the range of 60,001 to 300,000 (35.6% of clusters), while our proposal reduces that number to 123 (13.2%). In comparison, the number of

Land area intervals	Proposed geography		Currently used geography	
	All	> 1 municipalities	All	> 1 municipalities
< 1,000	773	602	478	84
1,000–2,000	96	96	47	39
2,000–3,000	36	36	15	15
3,000–4,000	20	20	16	16
4,000–5,000	3	3	8	8
5,000–6,000	3	3	4	4
6,000–7,000	0	0	5	5
7,000–8,000	0	0	1	1
8,000–9,000	0	0	1	1
9,000–10,000	0	0	2	2
10,000–11,000	0	0	2	2
11,000–12,000	0	0	2	2
12,000–13,000	0	0	4	4
13,000–14,000	0	0	1	1
> 14,000	0	0	1	1
Total	**931**	**760**	**587**	**185**

Tab. 1: Number of clusters by area intervals (km^2). Currently used and proposed geographies. Source: authors' results based on data from the Spanish Census of Population 2011 (Instituto Nacional de Estadística, INE).

Population intervals	Proposed geography		Currently used geography	
	All	> 1 municipalities	All	> 1 municipalities
< 20001	1	1	28	19
20,001–40,000	706	623	246	29
40,001–60,000	87	64	89	38
60,001–80,000	41	20	64	27
80,001–100,000	31	17	54	27
100,001–120,000	10	6	22	15
120,001–140,000	10	4	16	7
140,001–160,000	2	2	10	5
160,001–180,000	7	4	7	1
180,001–200,000	7	3	11	5
200,001–220,000	5	3	13	6
220,001–240,000	4	3	6	2
240,001–260,000	5	3	2	0
260,001–280,000	0	0	2	1
280,001–300,000	1	0	2	1
300,001–350,000	5	4	4	0
350,001–400,000	1	0	2	1
400,001–500,000	2	1	2	0
500,001–750,000	3	2	4	1
750,001–1,000,000	1	0	1	0
1,000,001–2,000,000	1	0	1	0
> 2,000,000	1	0	1	0
Total	**931**	**760**	**587**	**185**

Tab. 2: Number of clusters by population intervals. Currently used and proposed geographies. Source: authors' results based on data from the Spanish Census of Population 2011 (Instituto Nacional de Estadística, INE). Note: see footnotes 9 and 10 for group < 20,001.

clusters with the largest populations, which in most cases correspond to the main cities (already classified as single-municipality clusters), reveals few differences between both regionalisations.

6. Conclusions

This paper deals with a problem that is frequently seen when microdata from different statistical operations are made available for academic research and other uses: the lack of territorial detail derived from sampling or confidentiality restrictions. More specifically, microdata frequently refer to large units such as regions or provinces (NUTS 2 or NUTS 3 in the EU terminology) geographical levels that hamper detailed territorial analyses. When lower-level administrative units are included in the diffusion programmes, they are typically subject to a minimum population restriction. The microdata file associated with the Census of Population 2011 in the Spanish case exemplifies this situation: while the geography currently in use includes the specification of a local territorial reference in the case of municipalities over 20,000 inhabitants, the remaining municipalities are grouped within each province into a maximum of four clusters depending on the population interval to which they belong.

In this article we propose an approach in which a new geography is produced. This partition of the territory is designed to maximise the number of identified clusters of municipalities (so that the detail of the territorial reference used in the microdata file is increased), each of which exceeds a certain minimum population level, with the maximisation of the commuting links between the municipalities that constitute each cluster acting as a secondary objective. Such clusters are identified as subdivisions of a pre-existing set of LMAs (Martínez-Bernabeu et al., 2016). To achieve that goal, we have designed a new method based on a novel optimisation approach recently applied in the field of functional regionalisation, an evolutionary optimisation technique (GEA: Martínez-Bernabeu et al., 2012) previously used to define the LMAs. In this study, we have adapted the fitness function and the search operators of this technique to adapt to the objectives and restrictions of this specific problem.

The results, an application of the approach to the Spanish case, are designed to increase the territorial detail in the 2011 Census of Population microdata file to permit a more accurate analysis of the labour market at local levels. The resulting geography consists of 931 clusters of municipalities. Some of them (approximately 400) are roughly similar to the ones currently used (they basically correspond to municipalities exceeding the 20,000 inhabitants threshold). The rest (more than 500) are subdivisions of the 185 clusters that in the currently-used territorial division are formed by the aggregation of the municipalities with less than 20,000 inhabitants into four groups within each province. The new clusters are logically characterised by lower figures of both population and area, and allow for an increase in territorial resolution in the microdata file, while respecting the statistical constraints established by the National Institute of Statistics for the diffusion of individual data.

Since the new clusters have been conceived as subdivisions of the Spanish LMAs, this new regionalisation also permits an analysis at the level of LMAs (Martínez-Bernabeu et al., 2016), in contrast with the reference geography currently included in the microdata file, in which many of the clusters of municipalities are not contiguous and the diverse parts of the clusters are frequently separated by large distances. Moreover, pre-existing clusters have excessively (and unnecessarily) large areas and/or populations, and most of them consist of municipalities from different LMAs, to the detriment of an analysis of the interactions between and within LMAs. None of these drawbacks are present in the alternative regionalisation presented in this article. Moreover, if subsequent analyses find it useful, this subdivision of the territory into smaller clusters would allow for minor adjustments of the LMAs' boundaries. A forthcoming step in this research programme will involve the inclusion of the clusters' territorial codes in the Census 2011 microdata dataset for the seven variables listed previously, and its use in the analysis of commuting and migration behaviour at an individual level, as well as the analysis of the influence of the characteristics of the LMA/cluster of residence on the labour market outcomes of such individuals.

Finally, one incidental contribution of this article is the illustration of how the GEA (Martínez-Bernabeu et al., 2012) algorithm, originally designed for the delineation of LMAs, may be easily adapted to other related contexts through the modification of its fitness function and restrictions, according to the nature of the specific instance of regionalisation to which it is applied.

Acknowledgement

This work was supported by the Spanish Ministry of Economy and Competitiveness (grant number CSO2014-55780-C3-2-P, National R&D&i Plan 2013-2016). Census data were provided by the Instituto Nacional de Estadística (INE), an institution that is not responsible for the use of such data in this research.

References:

ALONSO, M. P., BEAMONTE, A., GARGALLO, P., SALVADOR, M. (2015): Local labour markets delineation: an approach based on evolutionary algorithms and classification methods. Journal of Applied Statistics, 42(5): 1043–1063.

BALAS, E., PADBERG, M. W. (1976): Set partitioning: A survey. SIAM review, 18(4): 710–760.

BARNHART, C., COHN, A. M., JOHNSON, E. L., KLABJAN, D., NEMHAUSER, G. L., VANCE, P. H. (2003): Airline crew scheduling. In Handbook of transportation science. New York, Springer US, 517–560.

CHAKRABORTY, A., BEAMONTE, M. A., GELFAND, A. E., ALONSO, M. P., GARGALLO, P., SALVADOR, M. (2013): Spatial interaction models with individual-level data for explaining labor flows and developing local labor markets. Computational Statistics and Data Analysis, 58: 292–307.

COCKINGS, S., HARBOOT, A., MARTIN, D., HORNBY, D. (2011): Maintaining existing zoning systems using automated zone-designed techniques: methods for creating the 2011 Census output geographies for England and Wales. Environment and Planning A, 43: 2399–2418.

COOMBES, M., BOND, S. (2008): Travel-To-Work Areas: the 2007 Review. London, Office for National Statistics.

COOMBES, M., GREEN, A., OPENSHAW, S. (1986): An efficient algorithm to generate official statistical reporting areas: The case of the 1984 travel-to-work areas revision in Britain. Journal of the Operational Research Society, 37: 943–953.

FALKENAUER, E. (1998): Genetic Algorithms and Grouping Problems. New York: John Wiley & Sons.

FLÓREZ-REVUELTA, F., CASADO-DÍAZ, J. M., MARTÍNEZ-BERNABEU, L. (2008): An evolutionary approach to the delineation of functional areas based on travel-to-work flows. International Journal of Automation and Computing, 5(1): 10–21.

FORTUNATO, S., BARTHELEMY, M. (2007): Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104: 36–41.

GOLDBERG, D. E. (1989): Genetic Algorithms in Search, Optimization, and Machine Learning. Addison, Wesley.

ISTAT (1997): I Sistemi Locali Del Lavoro 1991. Rome, Istituto Nazionale di Statistica.

ISTAT (2005): I Sistemi Locali Del Lavoro 2001. Rome, Istituto Nazionale di Statistica.

ISTAT (2014): Sistemi Locali Del Lavoro. Nota metodologica. Rome, Instituto Nazionale di Statistica [online]. Available at: http://www.istat.it/it/files/2014/12/nota-metodologica_SLL2011_rev20150205.pdf

LANCICHINETTI, A., FORTUNATO, S. (2011): Limits of modularity maximization in community detection. Physical Review E, 84: 066122.

LEVINE, D. (1996): A parallel genetic algorithm for the set partitioning problem. New York, Springer US, 23–35.

LAPORTE, G. (1992): The vehicle routing problem: An overview of exact and approximate algorithms. European Journal of Operational Research, 59(3): 345–358.

MARTIN, D. (2000): Towards the geographies of the 2001 UK Census of Population. Transactions of the Institute of British Geographers, New Series, 25: 321–332.

MARTÍNEZ-BERNABEU, L., CASADO-DÍAZ, J. M., FLÓREZ-REVUELTA, F. (2016): Improving the objective and constraint functions in optimisation-based functional regionalisation methods. Mimeo.

MARTÍNEZ-BERNABEU, L., FLÓREZ-REVUELTA, F., CASADO-DÍAZ, J. M. (2012): Grouping genetic operators for the delineation of functional areas based on spatial interaction. Expert Systems with Applications, 39(8): 6754–6766.

NEWMAN, M. E. J., GIRVAN, M. (2004): Finding and Evaluating Community Structure in Networks. Physical Review E, 69: 026113. doi: http://dx.doi.org/10.1103/PhysRevE.69.026113

ONS (2015): Methodology note on 2011 Travel to Work Areas. London: Office for National Statistics [online]. Available at http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/other/travel-to-work-areas/index.html

OPENSHAW, S. (1977a): A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling, Transactions of the Institute of British Geographers, New Series, 2: 459–472.

OPENSHAW, S. (1977b): Algorithm 3: a procedure to generate pseudo-random aggregations of N zones into M zones, where M is less than N, Environment and Planning A, 9: 1423–1428

OPENSHAW, S., RAO, L. (1995): Algorithms for re-engineering 1991 Census geography, Environment and Planning A, 27: 425–446.

SMART, M. W. (1974): Labour market areas: uses and definition. Progress in Planning, 69: 238–353.

TOTH, P., VIGO, D. (2001): The vehicle routing problem. Philadelphia, US, Society for Industrial and Applied Mathematics (SIAM).

Please cite this article as:

MARTÍNEZ-BERNABEU, L., CASADO-DÍAZ, J. M. (2016): Delineating zones to increase geographical detail in individual response data files: An application to the Spanish 2011 Census of population. Moravian Geographical Reports, 24(2): 26–36. Doi: 10.1515/mgr-2016-0008.